DOCUMENT RESUME

ED 441 034 TM 030 833

AUTHOR	Rudner, Lawrence M.
TITLE	Informed Test Component Weighting.
SPONS AGENCY	Maryland State Dept. of Education, Baltimore.
PUB DATE	2000-03-20
NOTE	11p.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Reliability; *Scores; Test Construction; Test Items; Validity
IDENTIFIERS	*Weighting (Statistical)

ABSTRACT

Testing programs that report a single score based on multiple choice and performance components must face the issue of how to derive the component scores. This paper identifies and logically evaluates alternative component weighting methods. It then examines composite reliability and validity as a function of weights, component reliability, component validity, and the correlation of the components. Weighting can make a big difference when combining a highly reliable test, such as a lengthy multiple-choice test, with a less reliable test, such as a short constructed-response test. A rational process that identifies and considers trade-offs in determining weights is suggested. (Contains 1 figure and 12 references.) (Author/SED)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 034

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as
received from the person or organization
originating it.
Minor changes have been made to
improve reproduction quality.
Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

L. Rudner

1

Informed Test Component Weighting

Lawrence M. Rudner

Maryland Assessment Research
Center for Education Success

A paper commissioned by the Maryland State Department of Education.

March 20, 2000

Department of Measurement, Statistics and Evaluation
University of Maryland, College Park 20742

TM030833

BEST COPY AVAILABLE

Author-
~~suggested~~ abstract

Informed Test Component Weighting

Testing programs that report a single score based on multiple choice and performance components must face the issue of how to derive the composite scores. This paper identifies and logically evaluates alternative component weighting methods. It then examines composite reliability and validity as a function of weights, component reliability, component validity and the correlation of the components. Weighting can make a big difference when combining a highly reliable test, such as a lengthy multiple-choice test, with a less reliable test, such as a short constructed-response test. A rational process that identifies and considers trade-offs in determining weights is suggested.

Informed Test Component Weighting

Lawrence M. Rudner
Maryland Assessment Research Center for Education Success &
ERIC Clearinghouse on Assessment and Evaluation

In many assessment situations, multiple tests or subtests are administered and the results are combined to form a single composite score. For example, today's commercially available achievement tests report a composite reading score formed by combining literal comprehension with inferential comprehension components. Other common uses of composite scoring can be found in combining the verbal and mathematical reasoning ability components of the SAT to form a composite score, interest inventories and personality tests, which often combine tests of different traits and in employment decision-making which is generally based on a combination of factors. Today, many large scale assessment programs are combining components that incorporate traditional multiple-choice items with components that incorporate newer performance items.

The manner in which composite scores are formed raises a variety of methodological and policy issues. For example, do we add more weight to a longer but less reliable performance assessment? Or should we add less weight? This problem is not new and a host of methods have been proposed. A half of a century ago, Gulliksen (1950) devoted a 50 page chapter to the topic. Thirty years ago, Wang and Stanley (1970) wrote a comprehensive 40 page review. Dawes (1976) revised the problem in his classic article on equally weighted measures. More recently Wainer and Thissen (1993) discussed the issue in the context of combining multiple-choice and constructed-response tests. Most of the literature suggests that weighting doesn't matter. Weighting does not appear to reduce overall error and the results of different weighting methods are often highly correlated. The effect of weights on the validity of the composite score, however, has not been adequately addressed.

This paper identifies and logically evaluates alternative methods for weighting tests. Formulas are presented for composite reliability and validity as a function of component weights. Concluding that weighting can make a big difference when combining a highly reliable test, such as a lengthy multiple-choice test, with a less reliable test, such as a short constructed-response test, a rational process that identifies and considers trade-offs in determining weights is suggested.

Component Weighting Methods

As Gulliksen (1950) points out and Wainer and Thissen (1993) underscore

> *It should be noted that it is not possible to dodge the weighting problem if any decisions are to be made. Occasionally, we hear the suggestion that scores simply be added*

together without bothering about the problem of weighting. No matter what scores we add, the problem is not avoided. (p 312)

Weighting methods are either implicit or explicit. This section identifies and briefly evaluates some of the more appealing approaches. The interested reader is referred to Wang and Stanley (1970) and Gulliksen (1950) for a more thorough reviews of non-IRT methods.

Implicit Approaches

Adding raw scores - Perhaps the simplest approach to combining test components is to simply add together the total number of correct responses. If the tests are developed following a blue print, then the number of items within each domain should be a fair representation of the domain's relative importance. In theory, by adding raw scores, a 100 item test would carry twice the weight of a 50 item test when the raw scores are combined. The logic, however, is faulty. Suppose you have an easy 100 item test with an extremely small variance and a moderately difficult 50 item test with a large variance. In this example, the 100 item test is like adding a constant to the 50 item score and contributes little to the variance of the composite scores. The 100 item test is not weighted twice that of the 50 item test. The effective weight will be proportional to the component variance. Another issue is that equal item weighting fails to consider differences in item importance. A lengthy algebra solution cannot be considered equal to recognizing an inequality on a multiple-choice test.

IRT Modeling - Rather than tackling the issue, one could simultaneously calibrate the items across all components and use an IRT model to estimate each examinees ability on the composite scale. A logical inconsistency arises, however. In order to incorporate most of today's operational IRT models, one must assume the composite is unidimensional. If the construct is unidimensional, then one should not be using the less efficient constructed-response items. If the construct is not unidimensional, then one should not be using IRT as IRT models do not appear to be robust to violations of the unidimensionality assumption (Dawadi, 1999; Harrison, 1986). Deriving theta from simultaneously calibrated one parameter IRT items is equivalent to summing the item scores. Deriving theta from a two parameter model is equivalent to weighting by the discrimination of the items within each component. The three parameter model would be influenced by both the discrimination parameter and the theta estimate.

Explicit Approaches

Weight by Difficulty - Instructors often weight items or sections of classroom tests based on their feel for the task difficulty. The same concept can be applied with empirical data. The approach appears to be attractive as it provides additional reward for mastering particularly difficult concepts. However, the converse is also true. The method punishes students more severely for missing these more difficult items. Weighting by easiness just reverses the penalties.

Reliability Weighting - Giving more reliable components heavier weights is intuitively appealing. The error associated with the composite score would be less if the more reliable measure were more heavily weighted. First, there is a problem of operationalizing these weights.

As seen in equation (1) below, weighting by component reliability will not maximize composite reliability. Second, as we will argue later, maximizing reliability is not necessarily a worthwhile goal. If that is one's goal, then optimal weights can be determined using Monte-Carlo techniques and equation (1) or by setting the first derivative of (1) with respect to w_1/w_2 to zero and solving.

Validity Weighting - A variety of methods can be used to maximize the validity of the composite scores. Multiple regression provides component weights that maximize the correlation of the composite with an external criterion. There is the well documented issue of shrinkage. The regression optimizes weights for the given data set. The weights may be less than optimal for other datasets. The validity coefficients themselves could be used as weights. This, however, would be even less optimal as it fails to consider the intercorrelations among the predictors. Further, determining a criterion in order to estimate validity coefficients is not always straightforward. As with maximizing reliability, maximizing validity may not be the most desirable goal.

Formulas for Composite Reliability and Validity

Let random variables $\underline{X} = [X_1, X_2]$ denote two components and random variable Y denote scores on a criterion variable. Further let $L = w_1X_1 + w_2X_2$ denote the weighted composite test score. To simplify calculations, set variances equal to unity. As a result $ó_1 = ó_2 = ó_y = 1.0$, $ñ_{12} = ó_{12}$ and $ñ_{Yi} = ó_{Yi}$ for i=1,2.

Wang and Stanley (1970, p. 672) provide a general formula for the reliability of a composite L composed of n variables. Solving for n=2 variables and simplifying, we have the reliability of a composite as a function of the weights, component reliabilities and the positive correlation between the components.

$$\rho_{LL'} = \frac{w_1^2\rho_{11'} + w_2^2\rho_{22'} + 2w_1w_2\rho_{12}}{w_1^2 + w_2^2 + 2w_1w_2\rho_{12}} \qquad (1)$$

From (1), we can derive that:

a) *the lowest possible value for the composite reliability is the reliability of the less reliable component.*
b) *if the components are correlated then the composite reliability can be higher than the reliability of either component.*
c) *if the component reliabilities are the same, then the composite reliability is maximum when the weights are the same.*

Winer (1971, p105) provides an equation for the squared product moment correlation between criterion variable Y and L as the product of the variance-covariance matrix, the scalar array of weights, and the scalar array of component-criterion variable covariances. Solving for n=2

predictor variables and simplifying yields the multiple correlation of a composite with a criterion variance as a function of the weights, component validities and the correlation between the components.

$$\rho_{yL} = \frac{w_1\rho_{y1} + w_2\rho_{y2}}{\sqrt{w_1^2 + w_2^2 + 2w_1w_2\rho_{12}}} \qquad (2)$$

Linear regression provides the weights that maximize the correlation between a composite and a criterion. Thus, the square root of the Multiple R associated with linear regression provides the maximum value for (2).

$$\max\rho_{yL} = \sqrt{\frac{\rho_{y1}^2 + \rho_{y2}^2 - 2\rho_{y1}\rho_{y2}\rho_{12}}{1-\rho_{12}^2}} \qquad (3)$$

From (2) and (3), we can deduce

a) the lowest possible value for the composite validity is the validity of the less valid component

b) the composite validity can be higher than the validity of either component.

c) if the component validities are the same then the composite validity is maximum when the weights are the same.

d) the maximum possible composite validity increases as the component correlation decreases.

Examples

This section provides an example to illustrate the effect of component weights on the composite test reliability and composite test validity. The analysis is based on the Biology AP examination as reported by Wainer and Thissen (1993). Component 1 is comprised of multiple choice items with a reliability of .93. Component 2 is comprised of constructed response items with a reliability of .68. The correlation of the two is .73 (0.92 unattenuated).

From Table 1, one can see there is very little change in the reliability as the weights change from ∞ to 1:1. The reliability starts to drop precipitously as extra weight is given to the less reliable constructed-response component. Here, the inclusion of the constructed-response component hurts reliability, regardless of the weighting.

Table 1: Composite test reliability as a function of select weights

weight W_1/W_2	∞	8/1	4/1	2/1	1/1	1/2	1/4	1/8	0
reliability $\tilde{n}_{LL'}$	.93	.93	.94	.92	.89	.83	.77	.73	.68

Looking at reliability data from a large number of composite tests where the longer multiple-choice component had a higher reliability, Wainer and Thissen (1993) concluded that "*whatever is being measured by the constructed-response section is measured better by the multiple choice section. These seven tests are but a sample. We have never found any test .. for which this is not true.*" These powerful words will hold as long as both sections have comparable validity and the constructed response section has lower reliability.

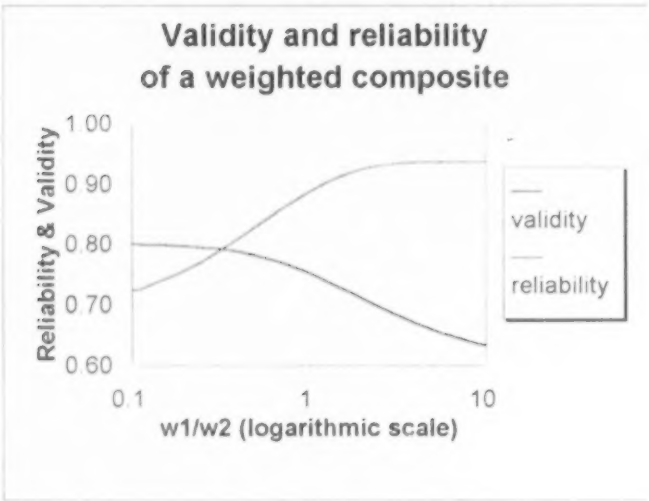
But, what if we value the content of the constructed-response more? That is, what if that section is more highly correlated with a criterion than the multiple choice section? McCornack (1956) discussed this over 40 years ago. The studies of his day mostly looked at the effect of weighting components with high reliability on the composite reliability and concluded that weighting doesn't matter. McCornack criticized these studies for not considering composite validity. The criticism also applies to Wainer and Thissen.

Differential component validity for the Biology AP examination is modeled in Table 2 and Figure 1. The correlation of the more reliable multiple-choice component with the criterion is set, for this example, to 0.60 (.62 unattenuated), while the validity coefficient for the less reliable constructed-response component is set to .80 (.97 unattenuated). From Figure 1, one can see that composite reliability rises steadily as more weight is given to the more reliable multiple-choice component and starts to level off when $w_1=1.5w_2$. Composite validity slowly drops as w_1 approaches $.7w_2$ and then starts to drop more precipitously. Clearly, in this example, the weights have a profound effect on validity and reliability.

Table 2: Component characteristics for Figure 1

	Validity	Reliability
Component 1	.60	.93
Component 2	.80	.68

$\tilde{n}_{12} = .73$



Contrary to the oft-quoted principle that square root of the reliability places an upper limit on validity, composite validity is increasing while composite reliability is decreasing in this example. This is not a mistake in formulas (1) and (2). That principle doesn't apply here as the components are not highly correlated. As astutely noted by Feldt (1997), the principle applies when the reduced reliability represents a similar, but shorter test. Here, by changing the weights, we have changed the essential character of the test; our test now measures something different, the true scores represent a different construct.

Discussion

We have argued that implicit weighting methods may not yield the desired results and that explicit weighting can seriously impact composite validity and composite reliability. Further, weighting can have unsatisfactory consequences. Maximizing reliability can lead to lower validity. Maximizing validity can lead to lower reliability.

The question remains, how does one weight two components? As argued by Kennedy and Walstad (1997), weighting should be a rational process evaluating contributions and the trade-offs. In Figure 1, for example, if one feels consistency is extremely important and that a validity coefficient of .75 is adequate, then a w_1/w_2 between 1.2 and 2.0 should be supported. Conversely, if one feels validity is more important and that a reliability of about .75 is adequate, then weighting component 2 more heavily with a w_1/w_2 of about .5 should be adopted. In both cases, the trade offs between reliability and validity can be rationally considered.

It should be noted from formulas (1) and (2) that the ends of the composite validity and composite reliability curves asymptotically approach the individual component validities and reliabilities. As the correlation between the components goes up, the curves becomes less peaked.

BEST COPY AVAILABLE

As the correlation approaches unity, one could simply maximize reliability. Thus, if the component validities are each satisfactory or the components are fairly intercorrelated, the different weights will not make much difference on composite validity. We suspect this will be the case with many large scale assessments that incorporate alternative item types. Nevertheless, one needs estimates of component validity, or a very high component correlation, before one can dismiss the effect of component validity on the validity of the composite scores.

The absence of a natural criterion variable does not mean data cannot be collected and used to help instruct the decision. Surrogate markers for clinically defined end-points (Prentice, 1989) are commonly used in medical research. Traditional approaches for conducting content validity studies based on ratings of item-objective congruence and relevance are applicable. A variety of new quantitative techniques for conducting content validity studies based on multidimensional scaling have been offered (Sireci, 1998). Value judgements can also be employed instead of the criterion measure, as long as the value judgement is a statement of worth and not perceived difficulty. Again, formulas (1) and (2) and a graph similar to Figure 1 can be used to help make the weighting decision policy.

This research was supported with funds from the Maryland State Department of Education. The views and opinions expressed in this paper are those of the author and do not necessarily reflect those of the Maryland State Department of Education. The author is indebted to William D. Schafer and James S. Roberts, Department of Measurement, Statistics and Evaluation at University of Maryland for their help and comments on earlier drafts of this paper.

References

Dawadi, B.R. (1999). Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assumption. Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

Dawes, R. (1976). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist*, 34(7), 571-582.

Feldt, L.S. (1997). Can Validity Rise When Reliability Declines? *Applied Measurement in Education*, 10(4), 377-387.

Gulliksen, H.O (1950). *Theory of Mental Tests*. New York: John Wiley and Sons, Inc.

Harrison, D.A. (1986). Robustness of Irt Parameter Estimation to Violations of the Unidimensionality Assumption. *Journal of Educational Statistics*, 11(2), 91-115.

Kennedy, P & W.B. Walstad (1997). Combining Multiple-Choice and Constructed Response Test Scores: An Economists View, *Applied Measurement in Education*, 10(4), 359-375.

McCornack, R.L. (1956). A criticism of studies comparing weighting methods. *Journal of Applied Psychology*, 40, 343-345.

Prentice, R.L. (1989). Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria. *Statistics in Medicine*, 8(4), 431-440.

Sireci, S.G. (1998). Gathering and Analyzing Content Validity Data. *Educational Assessment*; 5 (4), 299-321.

Wainer, H and D. Thissen (1993). Combining Multiple-Choice and Constructed Response Test Scores: Toward a Marxist Theory of Test Construction, *Applied Measurement in Education*, 6(2), 103-118.

Wang, M.D and J.C. Stanley (1970). Differential Weighting: a Review of Methods and Empirical Studies. *Review of Educational Research*, 40, 663-705.

Winer, B. J. (1971). *Statistical Principles in Experimental Designs.* New York: McGraw-Hill.